Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models

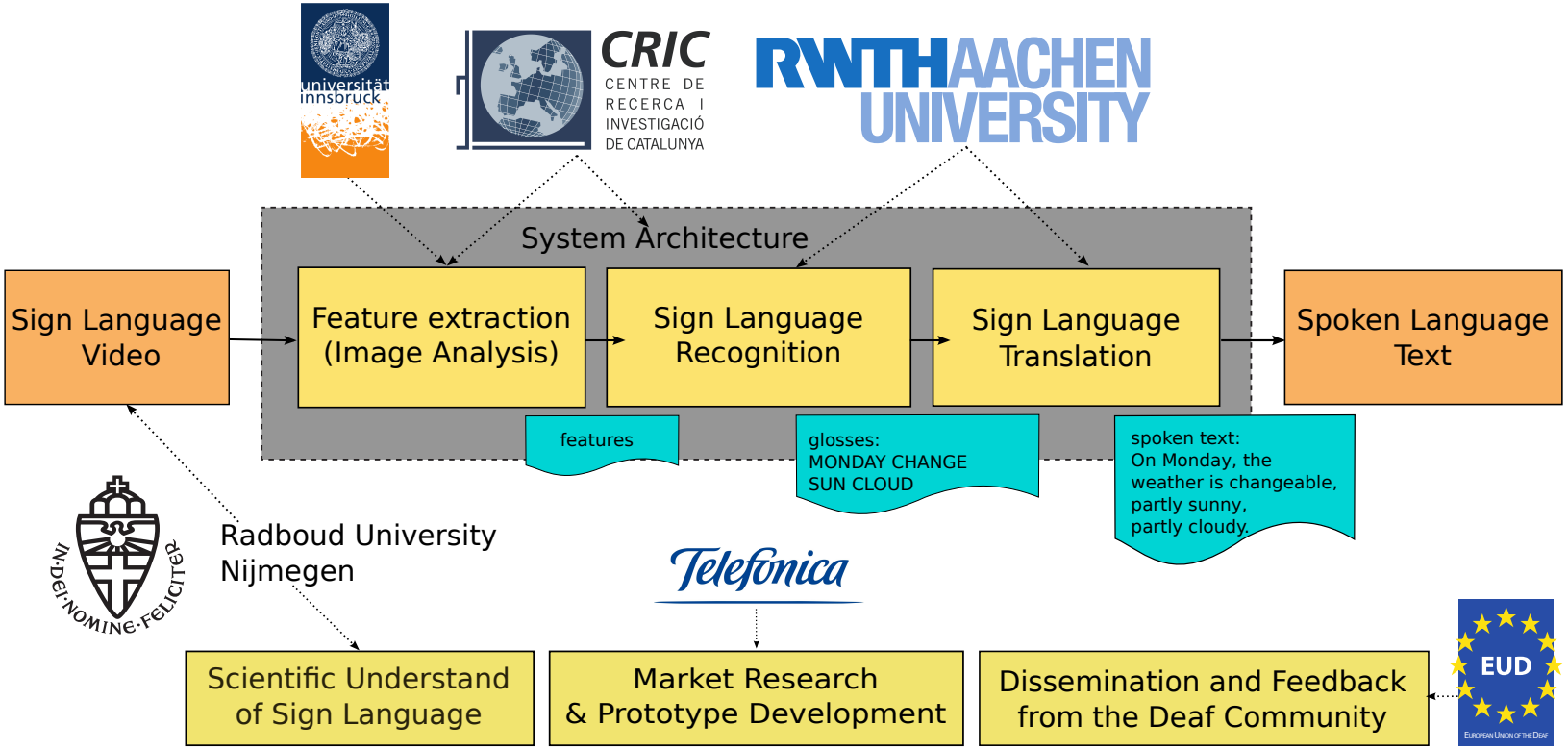
Christoph Schmidt, Oscar Koller, Hermann Ney [1]
Thomas Hoyoux, Justus Piater [2]

19.10.2013

[1] **Human Language Technology and Pattern Recognition Group**
Computer Science Department
RWTH Aachen University, Germany
{surname}@i6.informatik.rwth-aachen.de

[2] **Intelligent and Interactive Systems**
University of Innsbruck, Austria
{firstname}.{surname}@uibk.ac.at

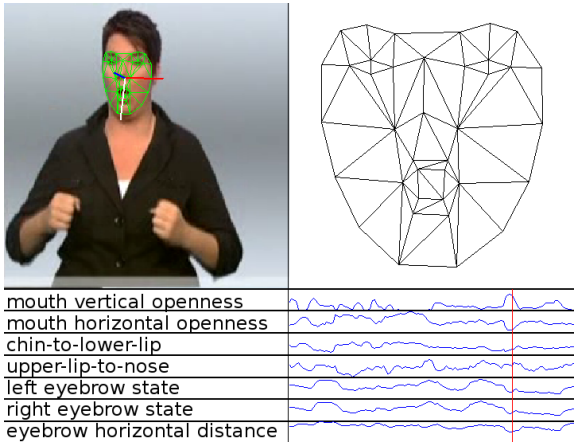
SignSpeak



- **Goal: translate a sign language video into a spoken language text**
- **Project Duration: April 2009-March 2012**

Active Appearance Models

- track salient points on the face
- extract high-level facial features:
  - mouth vertical openness
  - mouth horizontal openness
  - lower lip to chin distance
  - upper lip to nose distance
  - left eyebrow state
  - right eyebrow state
  - gap between eyebrows
- necessary: labeled data



RWTH-Phoenix-Weather Corpus

	DGS	German
signers	7	
editions	190	
duration[h]	3.25	
frames	293,077	
sentences	2,711	
glosses / words	17,744	33,190
vocabulary size	463	1,494
singletons	537	536

- video-based, large vocabulary corpus
- weather forecasts from public TV news, interpreted into DGS
- annotation: glosses, time boundaries on gloss level
- focus on hand-based features

Teaser:
new version with 645 editions coming soon at LREC 2014 !

Mouthing variants

ALPEN (“Alps”)

BERG (“mountain”)

- Some signs only differ in mouthing / mouth gestures
- Annotation of RWTH-Phoenix-Weather focused on hand-based features
- Manual refinement of annotation time consuming
- Idea: automatic refinement using feature extraction and clustering
- Avatar animation: use refined annotation to animate mouthings / facial expressions

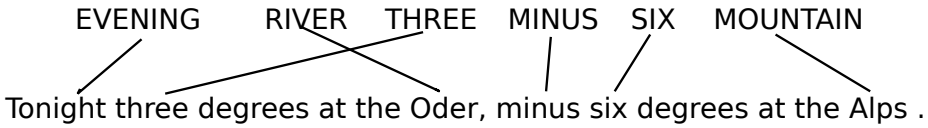
Clustering Approach

- Cluster variants using AAM features
- Use the context of the spoken language to drive the clustering
- For avatar animation: select representative video

- Define distance between two videos:
- Train Hidden Markov Model on one video
- Calculate Viterbi path of second video

Clustering Approach

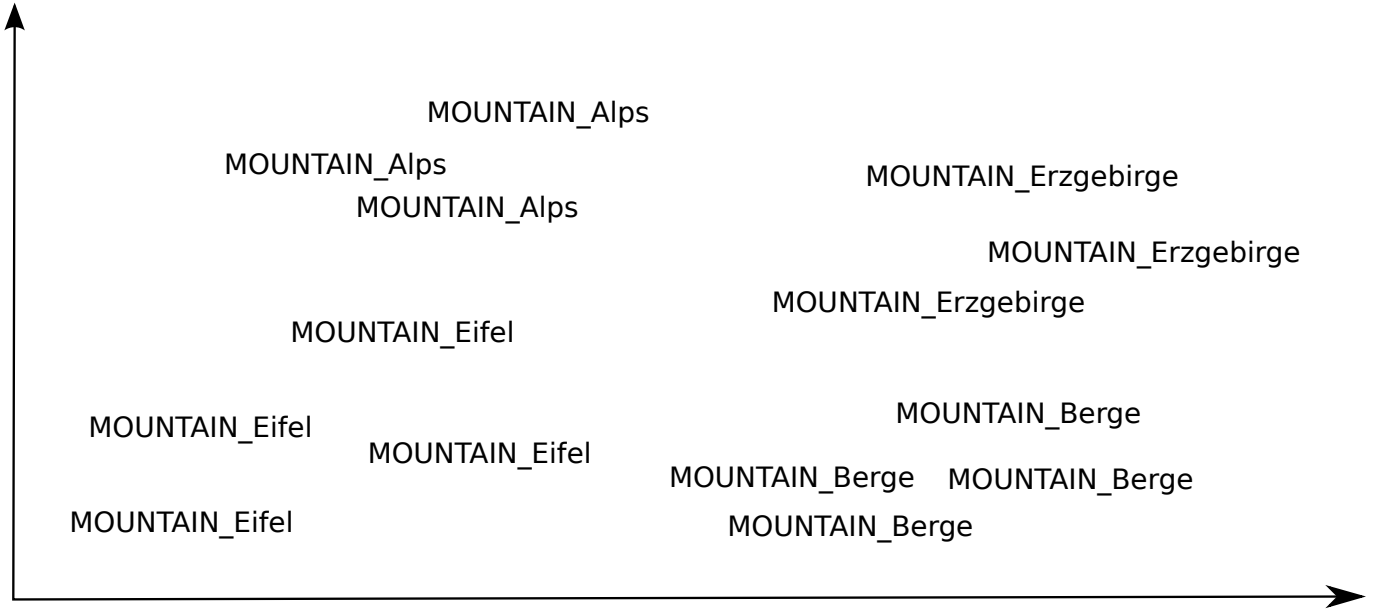
- **Align corpus**



- **Extract variants**

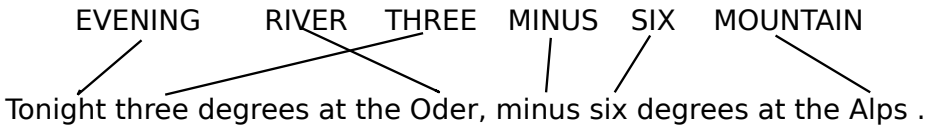
EVENING_tonight	EVENING_evening
RIVER_Oder	RIVER_Rhein
MOUNTAIN_Alps	MOUNTAIN_mountains

- **Cluster variants SL → Spoken**



Clustering Approach

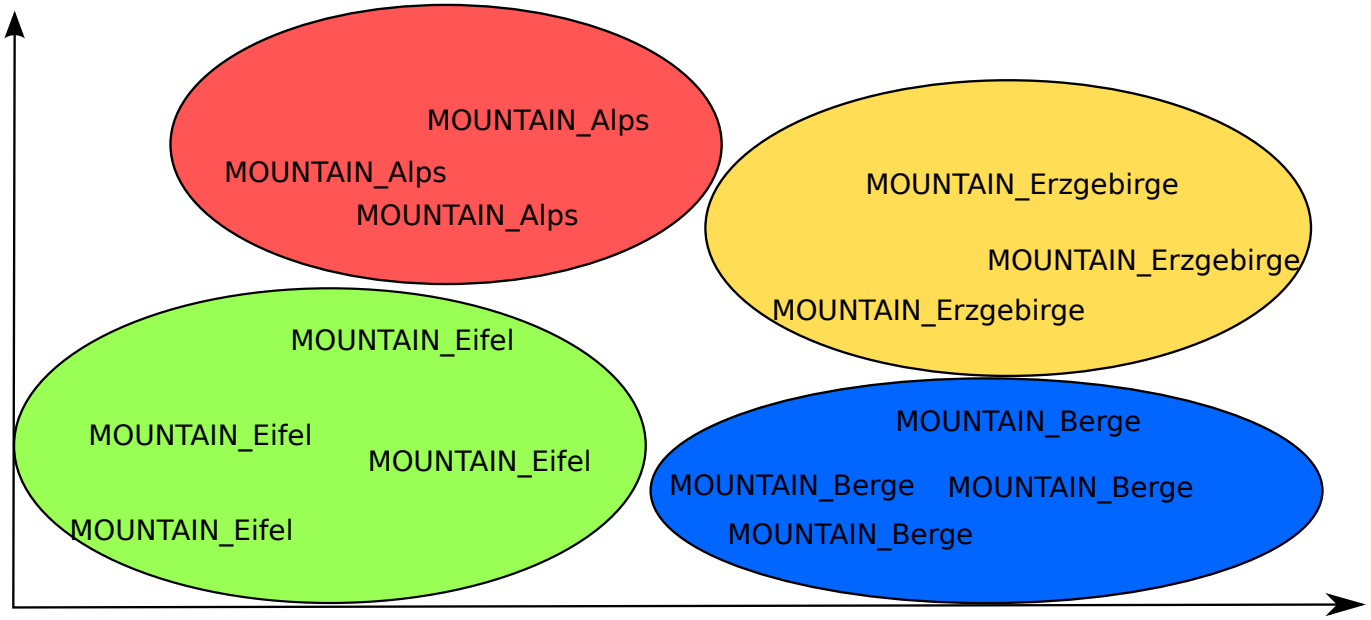
- **Align corpus**



- **Extract variants**

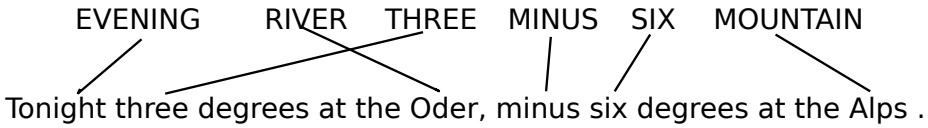
EVENING_tonight	EVENING_evening
RIVER_Oder	RIVER_Rhein
MOUNTAIN_Alps	MOUNTAIN_mountains

- **Cluster variants**
  SL → Spoken



Clustering Approach

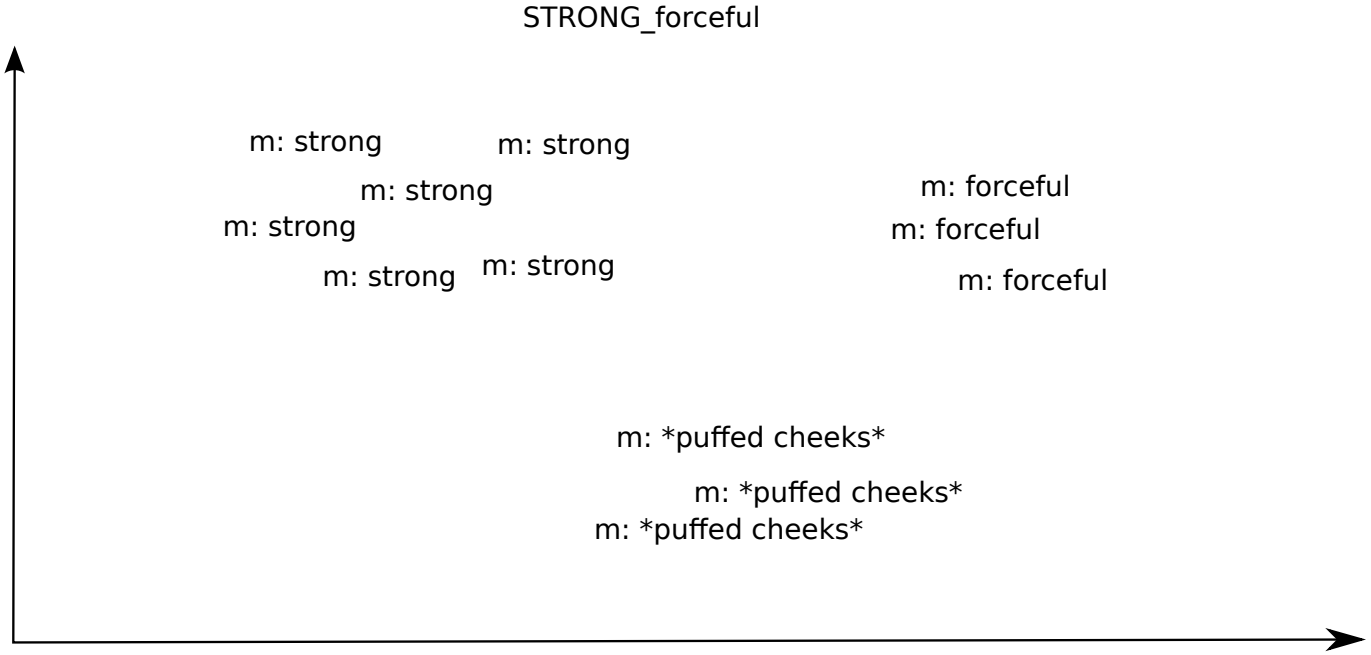
- **Align corpus**



- **Extract variants**

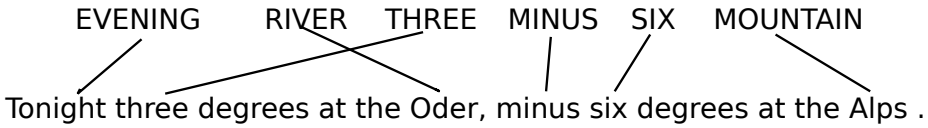
EVENING_tonight	EVENING_evening
RIVER_Oder	RIVER_Rhein
MOUNTAIN_Alps	MOUNTAIN_mountains

- **Cluster variants**
  Spoken → SL



Clustering Approach

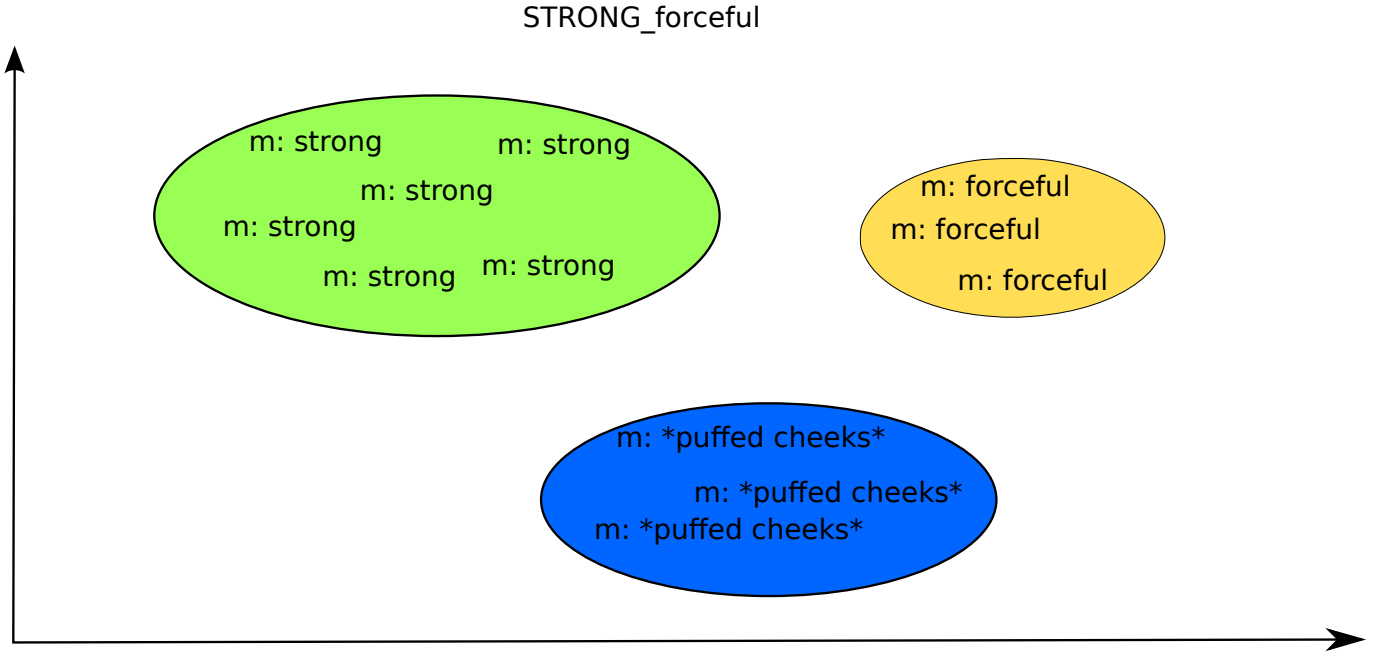
- **Align corpus**



- **Extract variants**

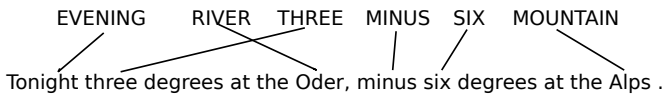
EVENING_tonight	EVENING_evening
RIVER_Oder	RIVER_Rhein
MOUNTAIN_Alps	MOUNTAIN_mountains

- **Cluster variants**
  Spoken → SL



Clustering Approach

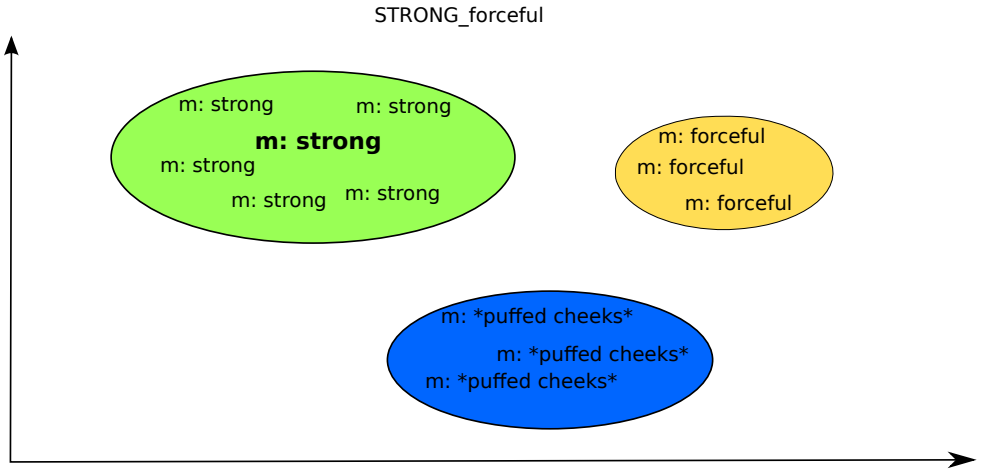
- **Align corpus**



- **Extract variants**

EVENING_tonight	EVENING_evening
RIVER_Oder	RIVER_Rhein
MOUNTAIN_Alps	MOUNTAIN_mountains

- **Cluster variants**
  Spoken → SL



- **Clustering algorithm: adaptive medoid-shift**
- **Select medoid of biggest cluster as representative video**

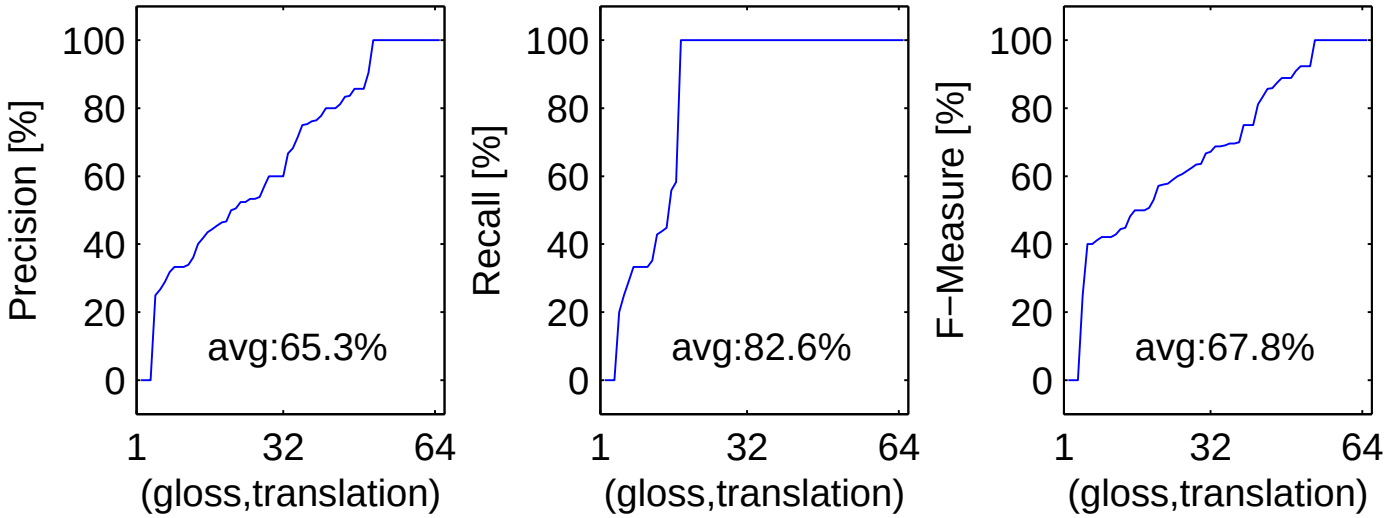
Experiments

- Annotate mouthings
  to evaluate clustering quality
- Select the most frequent glosses
  with more than one mouthing
- Select the most frequent contexts

GLOSS	context
MOUNTAIN	Alps
"	mountain
RIVER	Rhine
"	Oder
RAIN	rain
"	shower
EVENING	evening
"	night

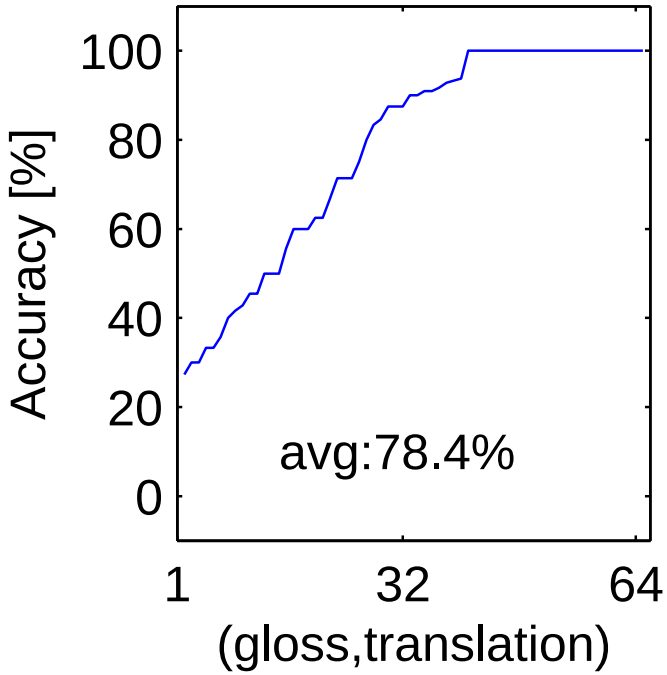
glosses	23
(gloss,translation) pairs	64
running glosses	640

Clustering results



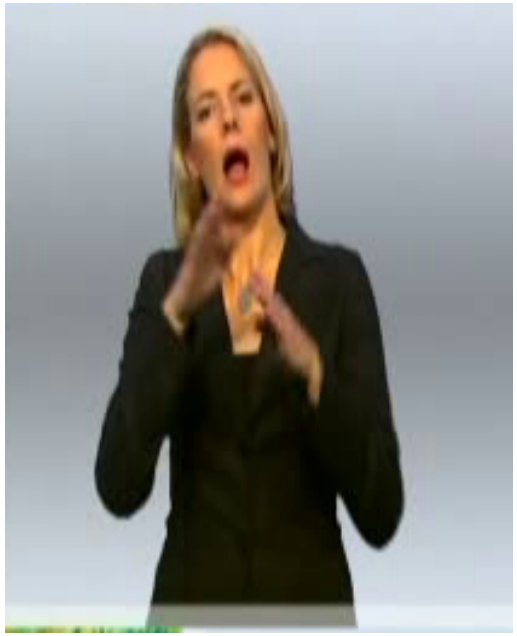
- **Precision: only same mouthings are in same cluster**
- **Recall: only different mouthings are in different clusters**
- **F-Measure: geometric mean of precision and recall**

Clustering results: biggest cluster



- **Accuracy: medoid has same mouthing as other cluster members**
- **The overall algorithm achieves accuracy of 78.4%**

Clustering results: Examples



- left video: (MOUNTAIN,Allgaeu)
- right video: medoid of biggest cluster
- Algorithm can recognize same mouthing even among different signers

Conclusion / Outlook

Conclusions:

- Clustering algorithm to detect variants in facial features
- Select representative video for avatar animation
- Achieves high accuracy

Outlook:

- improve low-level features: histogram of mouth area
- improve high-level features: HMM $\rightarrow$ visemes
- apply method beyond mouthings: facial expressions, head shake, etc.

Thank you for your attention

Christoph Schmidt

schmidt@i6.informatik.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/

Appendix: Annotated glosses

GLOSS	
BIT	NORTH
BUT	NOW
CALAMITY	RAIN
CAN	RIVER
COLD	SKY
COURSE	SNOW
DRY	SOUTH
ESPECIALLY	STRONG
EVENING	SUN
HIGH	TEMPERATURE
MORE	WIND
MOUNTAIN	

GLOSS	
ABEND	MEHR
ABER	NORD
BERG	REGEN
BESONDERS	SCHNEE
BISSCHEN	SONNE
FLUSS	STARK
GEWITTER	SUED
HIMMEL	TEMPERATUR
HOCH	TROCKEN
JETZT	VERLAUF
KALT	WIND
KOENNEN	

Appendix: Cluster Evaluation

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, \text{F-Measure} = \frac{2PR}{P + R}$$

- True Positive: same mouthings is in same cluster
- True Negative: different mouthings is in different cluster
- False Positive: different mouthings are in same cluster
- False Negative: same mouthings are in different cluster